
Laboratory 2:1 Microarray Data Processing for Affymetrix arrays

Day 2: Day August 17, 2004: 11:00 – 12:30

Ben Bolstad, Biostatistics, University of California, Berkeley

Key Concepts

- Computing Gene Expression measures for Affymetrix arrays
- Quality assessment of Affymetrix data

What you will be able to do at end of this section

- Perform basic inspection of raw Affymetrix array data
- Process raw data to produce
- Perform a basic quality assessment of the dataset

Introduction

This lab introduces you to some basic preprocessing tools for Affymetrix microarray data. We will make use of the BioConductor software packages *affy* and *affyPLM*.

First steps

Start R. At the command prompt type

```
library(affyPLM)
```

This will load the *affyPLM* and *affy* libraries and all of their dependencies. Next we will load the dataset we shall work on. Normally you'd read the CEL files which are outputted by the Affymetrix software, but for the purposes of this lab we shall use a dataset which has been previously read into the software. Type

```
data(Dilution)
```

this will load in the Dilution dataset. You can find out more about the Dilution dataset by typing

```
?Dilution
```

To get slightly more information about the data you should type

```
Dilution
```

Inspecting the data

The first stage in any microarray data analysis is to assess the data. First lets check the phenotypic data stored in the Dilution object. Do this by typing

```
pData(Dilution)
```

what do you observe?

Next lets explore the distribution of probe intensities for each chip. In particular type

```
boxplot(Dilution)
```

What do you observe?

Next examine a density plot of the Perfect Match intensities on the log2 scale

```
plotDensity(log2(pm(Dilution)), col=c("red", "pink", "blue", "green"), lty=1)
```

Lets look at the raw PM probe intensities for a few probesets. Type

```
pm(Dilution, "1001_at")
```

```
pm(Dilution, "924_s_at")
```

Now lets look at the probe-pattern for a random selection of probesets. Type

```
ourgenes <- sample(geneNames(Dilution), 6)
```

```
par(mfrow=c(2, 3))
```

```
for (i in 1:6){
```

```
  matplot(log2(pm(Dilution, ourgenes[i])), type="l", main=ourgenes[i])
```

```
}
```

What do you observe?

Computing expression measures

There are many ways to compute expression measures for affymetrix data. In this lab we will concentrate on the RMA expression measure. Type

```
eset <- rma(Dilution)
```

Lets inspect the expression values. Type

```
eset
```

For some summary information. Now type

```
exprs(eset)[1:5,]
```

to see a some of the numeric values for the expression values. Next lets explore the distribution of expression values by array. Type

```
par(mfrow=c(1,1))
```

```
boxplot(eset)
```

```
plotDensity(exprs(eset), col=c("red", "pink", "blue", "green"), lty=1)
```

How do these compare to what you saw with the raw data?

Some quality assessment

Normally you would use the fitPLM function to fit the models required for quality assessment. Because this is a computationally intensive procedure we will use another function which is faster, but produces slightly more approximate results. Type

```
Pset <-
```

```
rmaPLM(Dilution, output.param=list(weights=TRUE, pseudo.SE=TRUE))
```

Now lets look at some pseudo-chip images of various quantities.

```
image(Pset, which=2)
```

```
image(Pset, which=2, type="resids")
```

```
image(Pset, which=2, type="pos.resids")
```

```
image(Pset, which=2, type="neg.resids")
```

What do you observe?

Now a boxplot of NUSE (Normalized Unscaled Standard Errors)

```
boxplot(Pset)
```

and an image of the RLE (Relative Log Expression)

```
Mbox(Pset)
```

Do you observe anything? Do you conclude anything about the over all quality of your data?

If you have time

Compare RMA expression measures with MAS5 expression measures. First compute the MAS 5 expression measures

```
eset2 <- mas5(Dilution)
```

this might take a while. Next boxplot the two different expression measures

```
par(mfrow=c(1,2))
```

```
boxplot(data.frame(exprs(eset)),main="RMA Expression values",ylim=c(0,16))
```

```
boxplot(data.frame(log2(exprs(eset2))),main="MAS 5 Expression values",ylim=c(0,16))
```

What do you observe?

Next lets look at some MA plots.

```
A.rma <- 0.5*(exprs(eset)[,1] + exprs(eset)[,2])
```

```
M.rma <- (exprs(eset)[,1] - exprs(eset)[,2])
```

```
A.mas <- 0.5*(log2(exprs(eset2)[,1]) + log2(exprs(eset2)[,2]))
```

```
M.mas <- log2(exprs(eset2)[,1]) - log2(exprs(eset2)[,2])
```

```
par(mfrow=c(2,1))
```

```
ma.plot(A.rma,M.rma,ylim=c(-4,4), main="RMA")
```

```
ma.plot(A.mas,M.mas,ylim=c(-4,4),main="MAS 5")
```

What do you observe?

Final words

There is of course much more to the low level analysis of Affymetrix data than discussed here. The papers mentioned below are a good place to start from.

Appendix

1. Resources

i) Original Papers

- Bolstad, B.M., Irizarry R. A., Astrand, M., and Speed, T.P. (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* 19(2):185-193
- Rafael. A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs and Terence P. Speed (2003), Summaries of Affymetrix GeneChip probe level data *Nucleic Acids Research* 31(4):e15
- Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP (2003) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* .Vol. 4, Number 2: 249-264

ii) Software

- Bioconductor www.bioconductor.org
- RMAExpress www.stat.berkeley.edu/~bolstad/RMAExpress/RMAExpress.html

iii) Web Sites:

- www.stat.berkeley.edu/~bolstad/PLMImageGallery
- www.affymetrix.com

2. Demonstration Overheads

© 2004 Canadian Genetic Diseases Network